New Mexico CONSORTIUM

Los Alamos NATIONAL LABORATORY EST. 1943

Ultrascale Systems Research Center

a collaboration with the New Mexico Consortium and Los Alamos

Many unsolved problems are standing in the way of achieving exascale.

Scaling, systems software, and applications, to millions of nodes, and billion way parallelism, reliability, I/O, and storage, are research challenges at these scales.

USRC was created to address these challenges through collaboration.

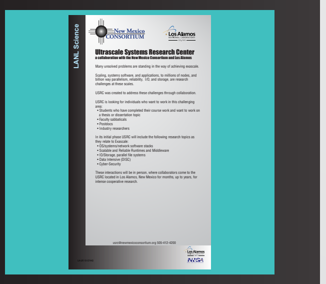
USRC is looking for individuals who want to work in this challenging area:

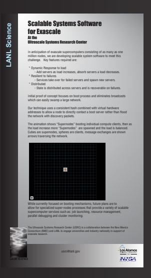
- Students who have completed their course work and want to work on a thesis or dissertation topic
- Faculty sabbaticals
- Postdocs
- Industry researchers

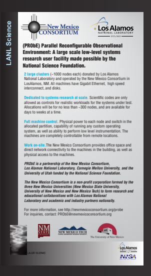
In its initial phase USRC will include the following research topics as they relate to Exascale:

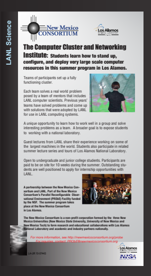
- OS/systems/network software stacks
- Scalable and Reliable Runtimes and Middleware
- IO/Storage, parallel file systems
- Data Intensive (DISC)
- Cyber-Security

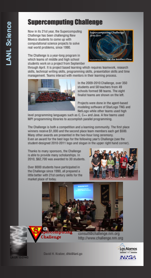
These interactions will be in person, where collaborators come to the USRC located in Los Alamos, New Mexico for months, up to years, for intense cooperative research.

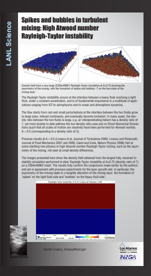


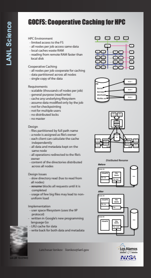


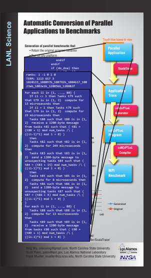








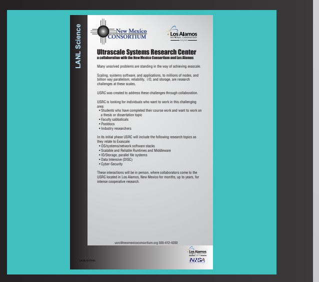


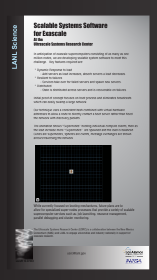


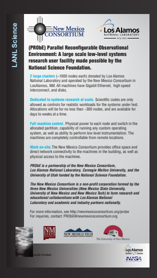
usrc@newmexicoconsortium.org 505-412-4200

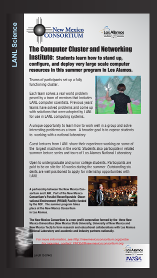
Los Alamos NATIONAL LABORATORY EST. 1943

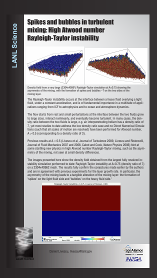
NNSA National Nuclear Security Administration

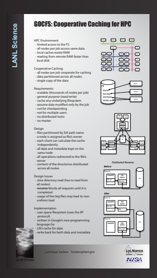


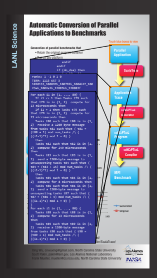












Scalable Systems Software for Exascale

At the
Ultrascale Systems Research Center

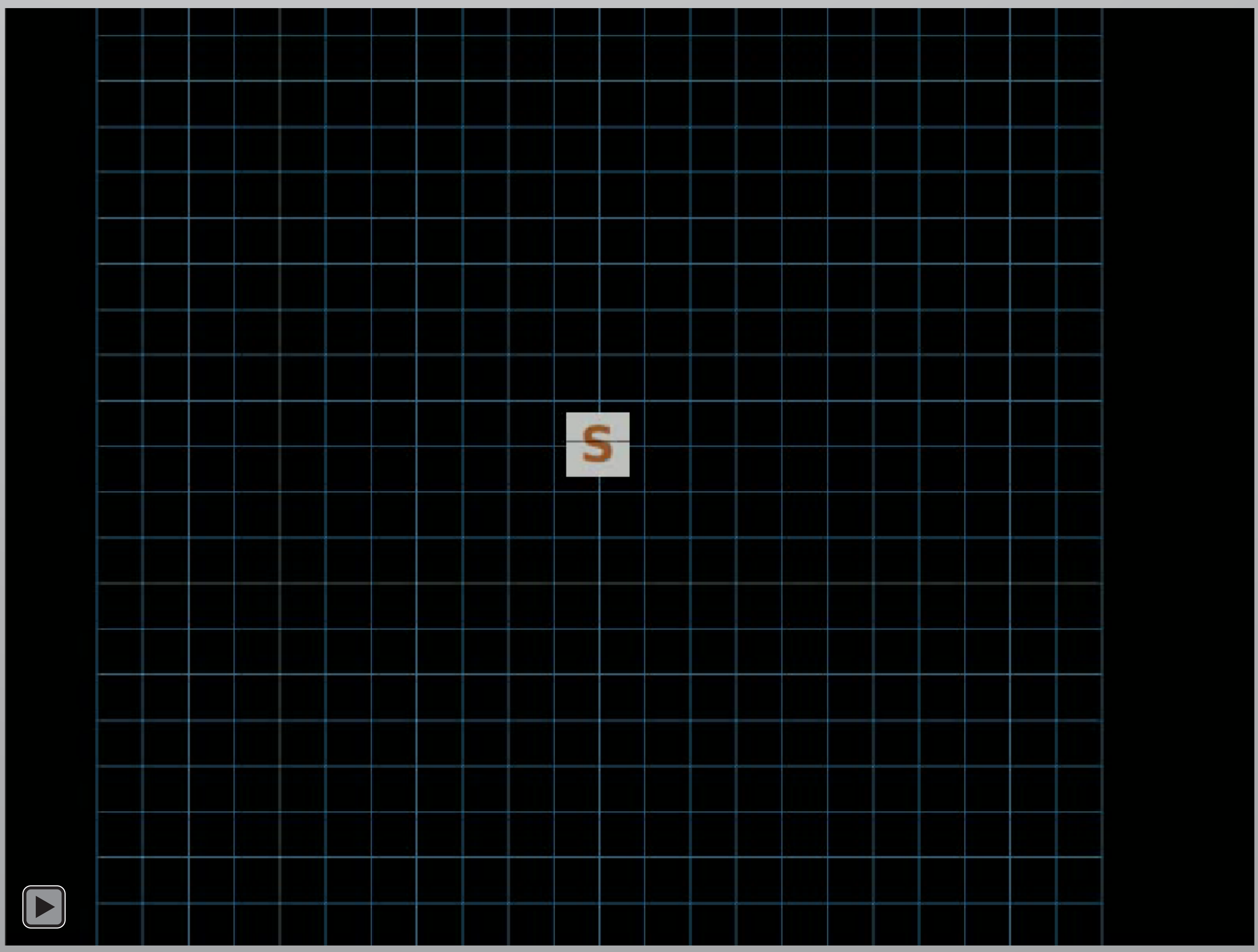
In anticipation of exascale supercomputers consisting of as many as one million nodes, we are developing scalable system software to meet this challenge. Key features required are:

* Dynamic Response to load
    - Add servers as load increases, absorb servers a load decreases.
* Resilient to failures
    - Services take over for failed servers and spawn new servers.
* Distributed
    - State is distributed across servers and is recoverable on failures.

Initial proof of concept focuses on boot process and eliminates broadcasts which can easily swamp a large network.

Our technique uses a consistent hash combined with virtual hardware addresses to allow a node to directly contact a boot server rather than flood the network with discovery packets.

The animation shows “Supernodes” booting individual compute clients, then as the load increase more “Supernodes” are spawned and the load is balanced. Cubes are supernodes, spheres are clients, message exchanges are shown arrows traversing the network.



While currently focused on booting mechanisms, future plans are to allow for specialized super-nodes processes that provide a variety of scalable supercomputer services such as: job launching, resource management, parallel debugging and cluster monitoring.

The Ultrascale Systems Research Center (USRC) is a collaboration between the New Mexico Consortium (NMC) and LANL to engage universities and industry nationally in support of exascale research.

New Mexico CONSORTIUM

Los Alamos
NATIONAL LABORATORY
EST.1943

(PRObE) Parallel Reconfigurable Observational Environment: A large scale low-level systems research user facility made possible by the National Science Foundation.

2 large clusters (~1000 nodes each) donated by Los Alamos National Laboratory and operated by the New Mexico Consortium in LosAlamos, NM. All machines have Gigabit Ethernet, high speed interconnect, and disks.

Dedicated to systems research at scale. Scientific codes are only allowed as controls for realistic workloads for the systems under test. Allocations will be for no less than ~300 nodes, and are available for days to weeks at a time.

Full machine control. Physical power to each node and switch in the allocated partition, capability of running any custom operating system, as well as ability to perform low level instrumentation. The machines are completely controllable from remote locations.

Work on-site. The New Mexico Consortium provides office space and direct network connectivity to the machines in the building, as well as physical access to the machines.

PRObE is a partnership of the New Mexico Consortium, Los Alamos National Laboratory, Carnegie Mellon University, and the University of Utah funded by the National Science Foundation.

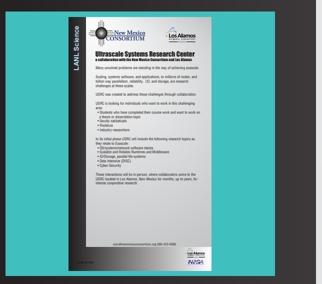
The New Mexico Consortium is a non-profit corporation formed by the three New Mexico Universities (New Mexico State University, University of New Mexico and New Mexico Tech) to form research and educational collaborations with Los Alamos National Laboratory and academic and industry partners nationally.

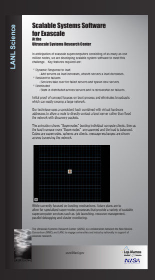
For more information, see http://newmexicoconsortium.org/probe
For inquiries, contact: PRObE@newmexicoconsortium.org

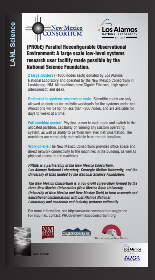
NM STATE UNIVERSITY

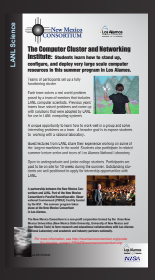
NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH • UNIVERSITY

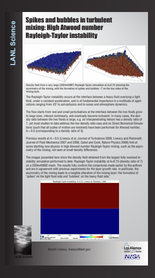
The University of New Mexico

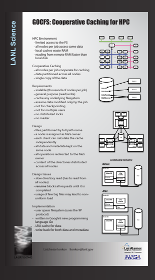


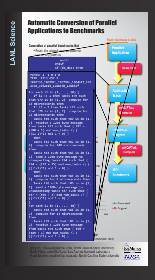








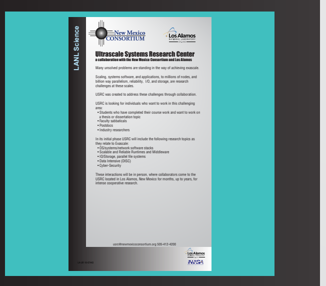
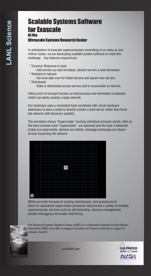
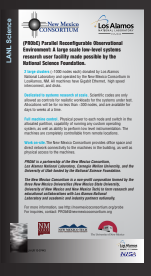
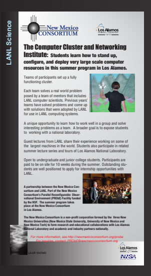
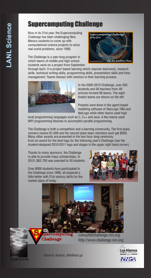
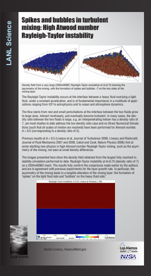
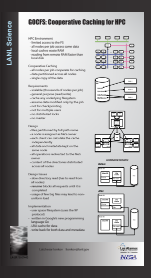
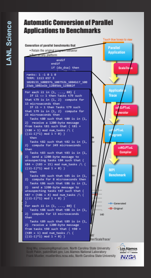




Los Alamos
NATIONAL LABORATORY
EST.1943

NNSA
National Nuclear Security Administration

LANL Science

New Mexico
CONSORTIUM

Los Alamos
NATIONAL LABORATORY
EST. 1943

The Computer Cluster and Networking Institute: Students learn how to stand up, configure, and deploy very large scale computer resources in this summer program in Los Alamos.

Teams of participants set up a fully functioning cluster.

Each team solves a real world problem posed by a team of mentors that includes LANL computer scientists. Previous years' teams have solved problems and come up with solutions that were adopted by LANL for use in LANL computing systems.

A unique opportunity to learn how to work well in a group and solve interesting problems as a team. A broader goal is to expose students to working with a national laboratory.

Guest lectures from LANL share their experience working on some of the largest machines in the world. Students also participate in related summer lecture series and tours of Los Alamos National Laboratory.

Open to undergraduate and junior college students. Participants are paid to be on site for 10 weeks during the summer. Outstanding students are well positioned to apply for internship opportunities with LANL.

A partnership between the New Mexico Consortium and LANL. Part of the New Mexico Consortium's Parallel Reconfigurable Observational Environment (PRObE) Facility funded by the NSF. The summer program takes place at the New Mexico Consortium in Los Alamos.

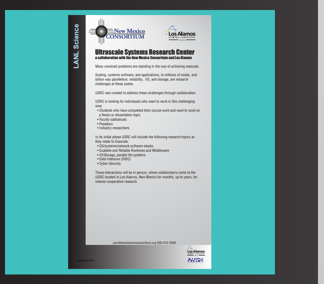
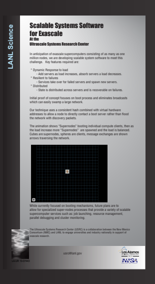
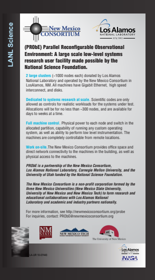
The New Mexico Consortium is a non-profit corporation formed by the three New Mexico Universities (New Mexico State University, University of New Mexico and New Mexico Tech) to form research and educational collaborations with Los Alamos National Laboratory and academic and industry partners nationally.

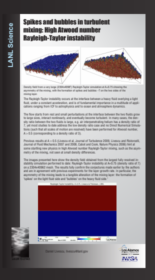
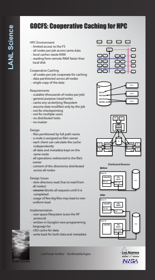
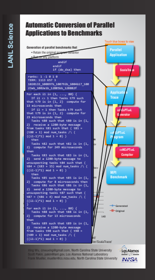
For more information, see http://newmexicoconsortium.org/probe
For inquiries, contact: PRObE@newmexicoconsortium.org

Los Alamos
NATIONAL LABORATORY
EST. 1943

NNSA
National Nuclear Security Administration

LANL Science

Supercomputing Challenge

Now in its 21st year, the Supercomputing Challenge has been challenging New Mexico students to come up with computational science projects to solve real world problems, since 1990.

The Challenge is a year-long program in which teams of middle and high school students work on a project from September through April. It is project based learning which requires teamwork, research skills, technical writing skills, programming skills, presentation skills and time management. Teams interact with mentors in their learning process.

In the 2009-2010 Challenge, over 350 students and 58 teachers from 49 schools formed 98 teams. The eight finalist teams are shown on the left.

Projects were done in the agent-based modeling software of StarLogo TNG and NetLogo while other teams used high level programming languages such as C, C++ and Java. A few teams used MPI programming libraries to accomplish parallel programming.

The Challenge is both a competition and a learning community. The first place winners receive $1,000 and the second place team members each get $500. Many other awards are presented in the two-hour long ceremony. Even an award for the best logo for the following year's Challenge (see the student-designed 2010-2011 logo and slogan in the upper right hand corner).

Thanks to many sponsors, the Challenge is able to provide many scholarships. In 2010, $62,700 was awarded to 30 students.

Over 8000 students have participated in the Challenge since 1990, all prepared a little better with 21st century skills for the market place of today.ċċċċċċċċċċ

Computational Modeling
Science Math
Supercomputing
Challenge

consult@challenge.nm.org
http://www.challenge.nm.org

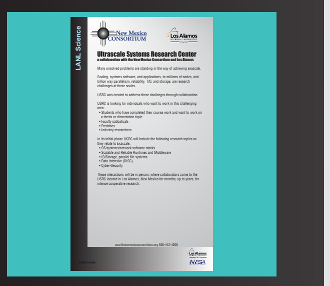
Los Alamos
NATIONAL LABORATORY
EST. 1943

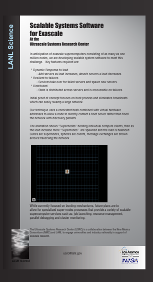
NNSA
National Nuclear Security Administration

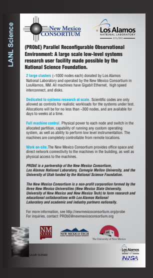
ċċċċċċċċċċċċċċċċċDavid H. Kratzer, dhk@lanl.goě

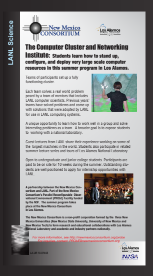
LA-UR 10-07443

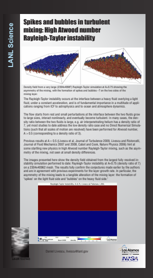
LANL Science

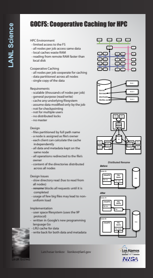


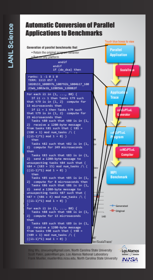




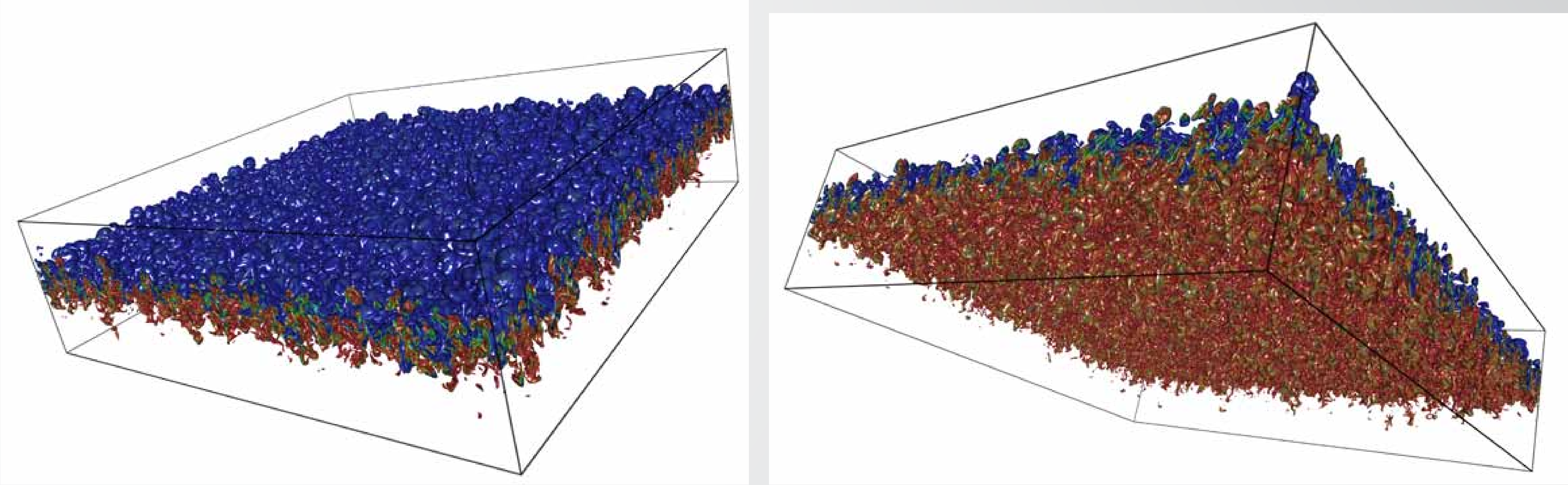








Spikes and bubbles in turbulent mixing: High Atwood number Rayleigh-Taylor instability



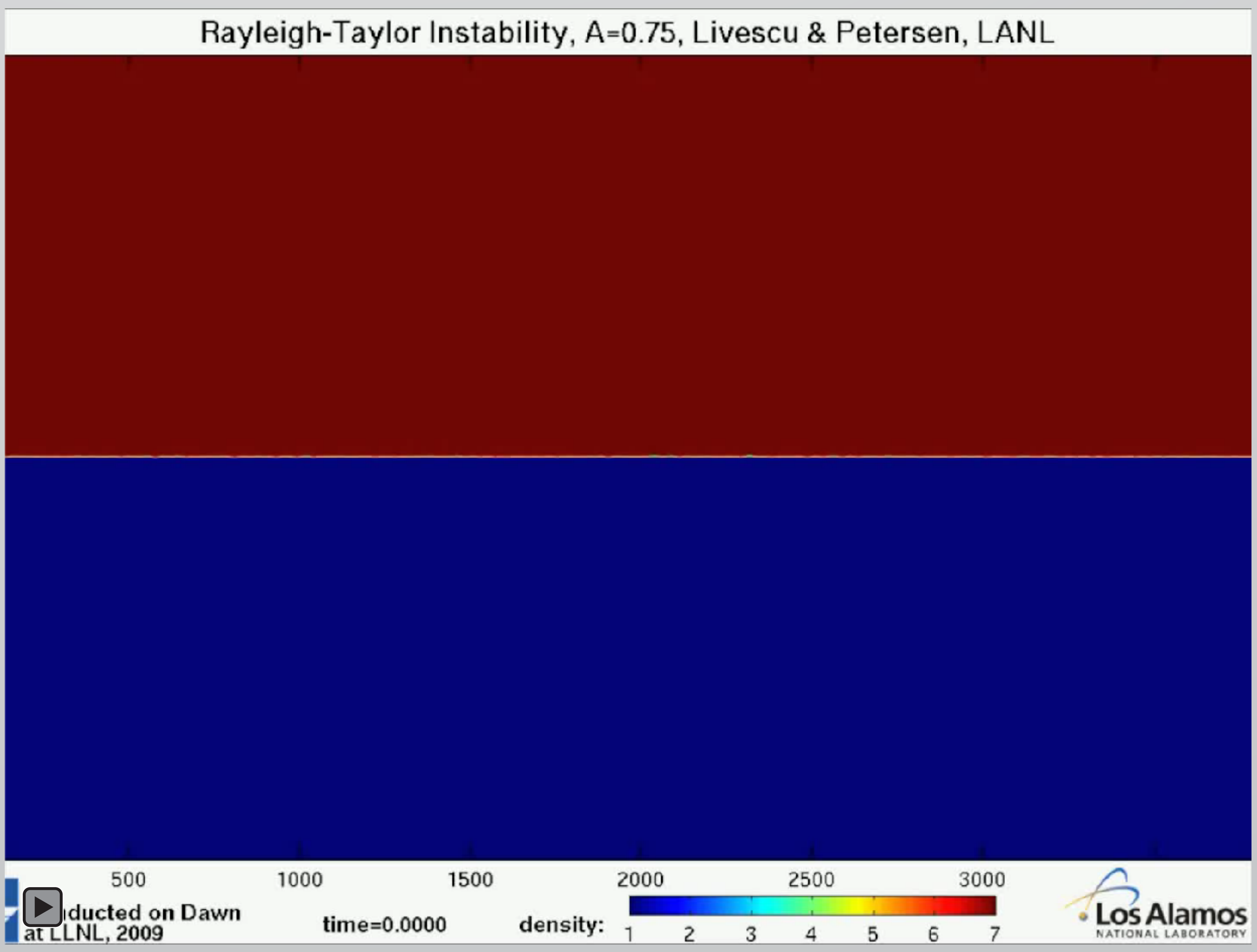
Density field from a very large (2304x4096^2) Rayleigh-Taylor simulation at A=0.75 showing the asymmetry of the mixing, with the formation of spikes and bubbles ~T on the two sides of the mixing layer.

The Rayleigh-Taylor instability occurs at the interface between a heavy fluid overlying a light fluid, under a constant acceleration, and is of fundamental importance in a multitude of applications ranging from ICF to astrophysics and to ocean and atmosphere dynamics.

The flow starts from rest and small perturbations at the interface between the two fluids grow to large sizes, interact nonlinearly, and eventually become turbulent. In many cases, the density ratio between the two fluids is large, e.g. air interpenetrating helium has a density ratio of 7, yet most studies to date address the low density ratio case and no Direct Numerical Simulations (such that all scales of motion are resolved) have been performed for Atwood number, A > 0.5 (corresponding to a density ratio of 3).

Previous results at A = 0.5 (Livescu et al, Journal of Turbulence 2009, Livescu and Ristorcelli, Journal of Fluid Mechanics 2007 and 2008, Cabot and Cook, Nature Physics 2006) hint at some startling new physics in high Atwood number Rayleigh-Taylor mixing, such as the asymmetry of the mixing, not seen at small density differences.

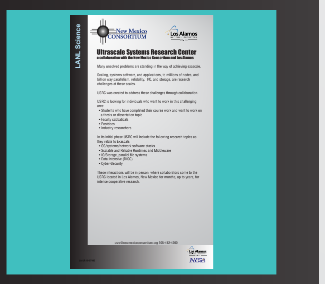
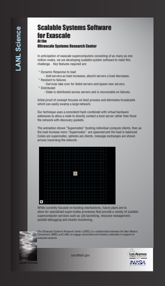
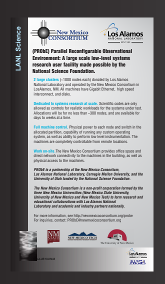
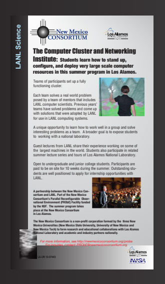
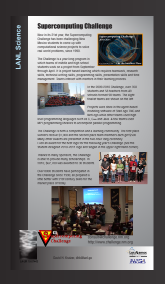
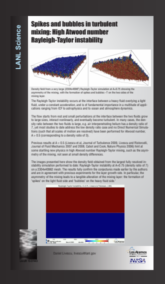
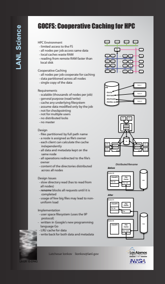
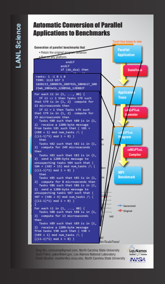
The images presented here show the density field obtained from the largest fully resolved instability simulation performed to date: Rayleigh-Taylor instability at A=0.75 (density ratio of 7) on a 2304x40962 mesh. The results fully confirm the conjectures made earlier by the authors and are in agreement with previous experiments for the layer growth rate. In particular, the asymmetry of the mixing leads to a tangible alteration of the mixing layer: the formation of "spikes" on the light fluid side and "bubbles" on the heavy fluid side."



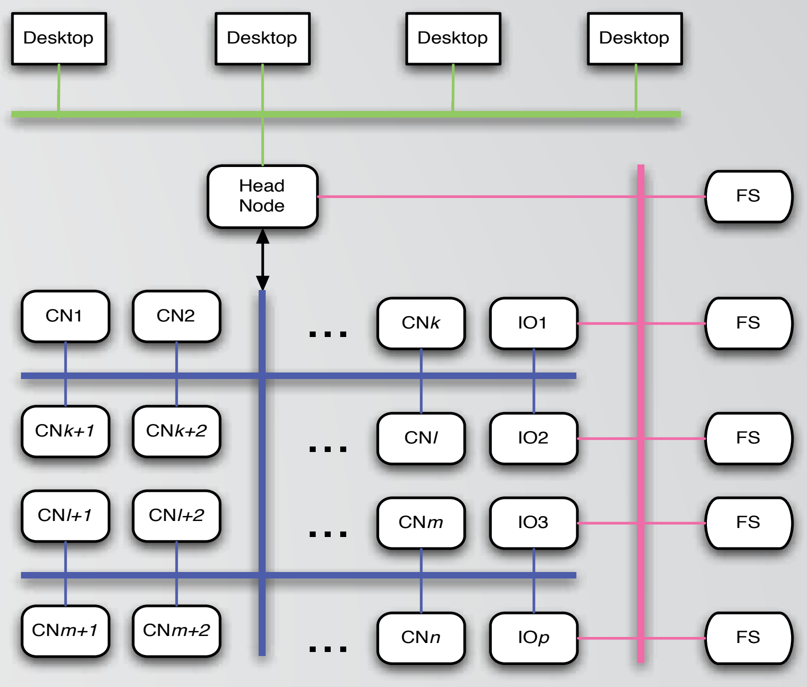
Daniel Livescu, livescu@lanl.gov

LA-UR 10-07443

LANL Science

GOCFS: Cooperative Caching for HPC

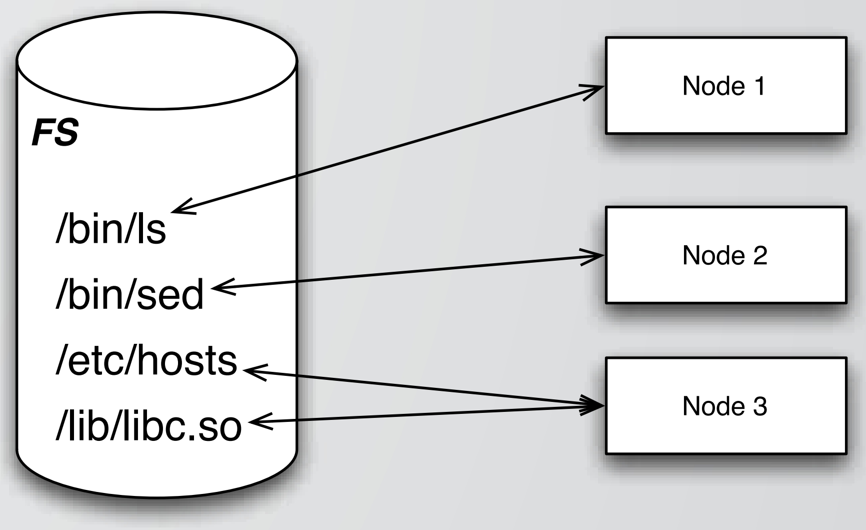


HPC Environment

- limited access to the FS
- all nodes per job access same data
- local caches waste RAM
- reading from remote RAM faster than local disk

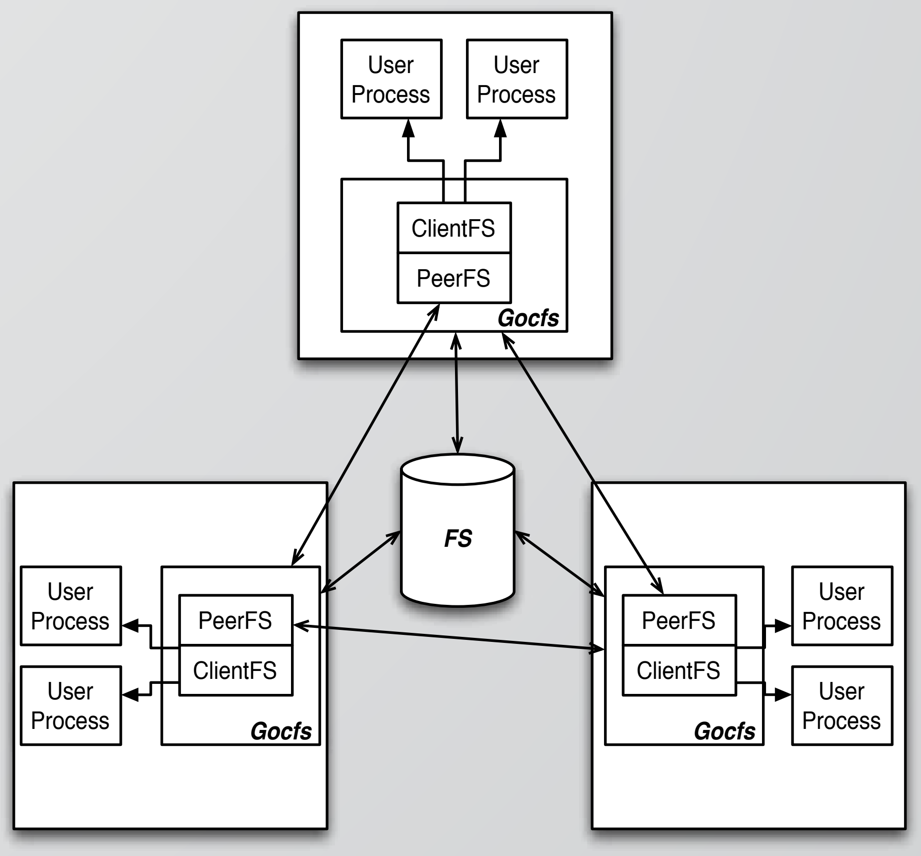
Cooperative Caching

- all nodes per job cooperate for caching
- data partitioned across all nodes
- single copy of the data



Requirements

- scalable (thousands of nodes per job)
- general purpose (read/write)
- cache any underlying filesystem
- assume data modified only by the job
- not for checkpointing
- not for multiple users
- no distributed locks
- no master



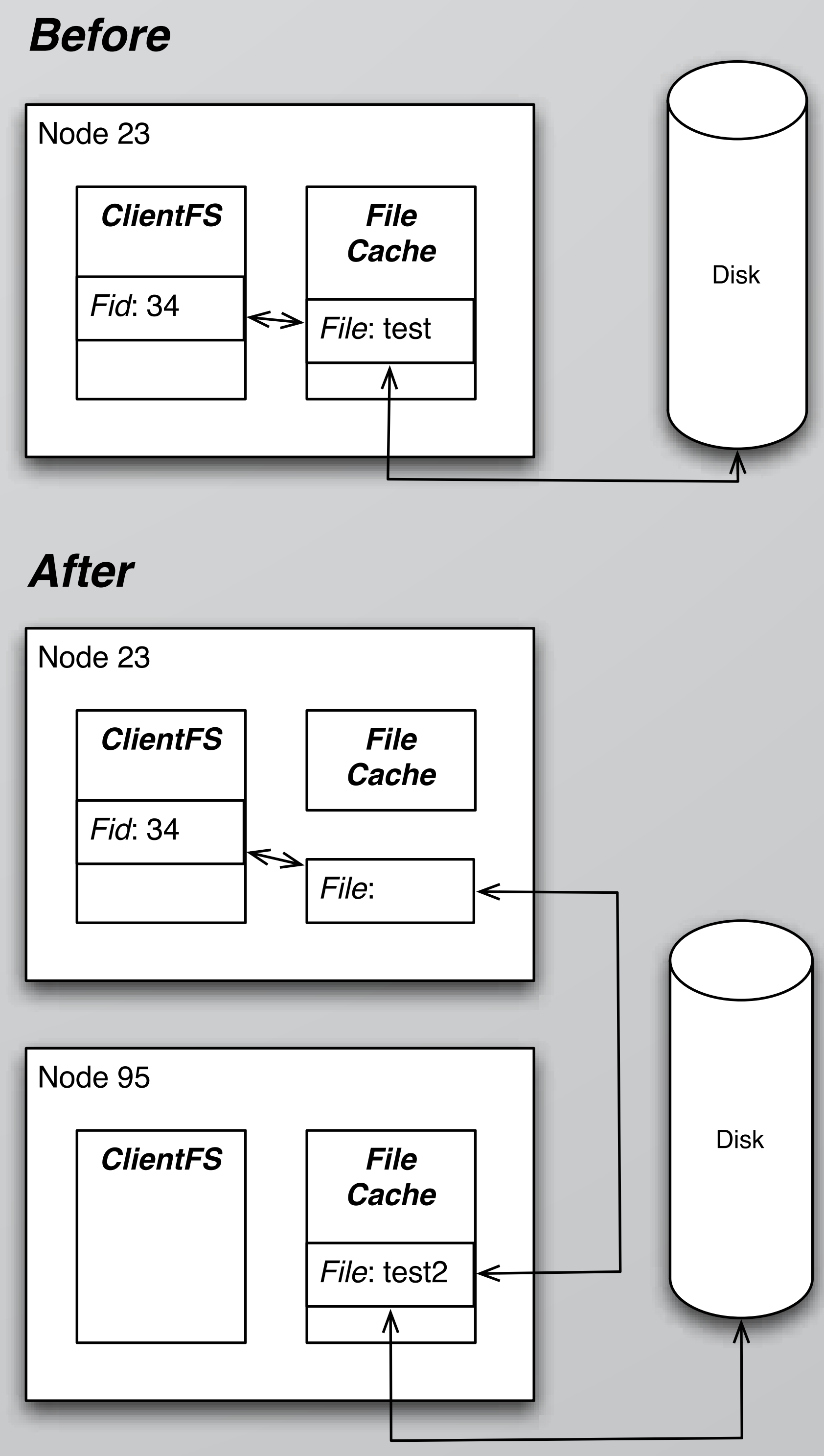
Design

- files partitioned by full path name
- a node is assigned as file's owner
- each client can calculate the cache independently
- all data and metadata kept on the same node
- all operations redirected to the file's owner
- content of the directories distributed across all nodes

Design Issues

- slow directory read (has to read from all nodes)
- ***rename*** blocks all requests until it is completed
- usage of few big files may lead to non-uniform load

Distibuted Rename



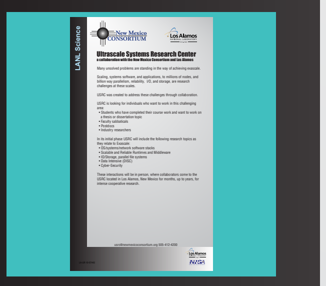
Implementation

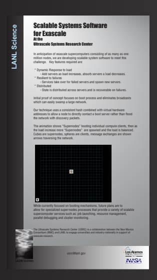
- user space filesystem (uses the 9P protocol)
- written in Google's new programming language Go
- LRU cache for data
- write back for both data and metadata

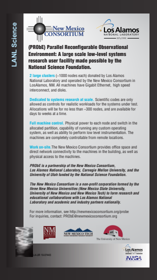
LA-UR 10-07443

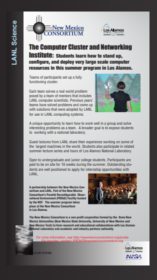
Latchesar Ionkov lionkov@lanl.gov

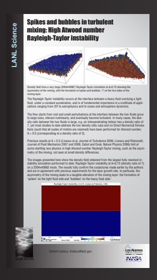
LANL Science

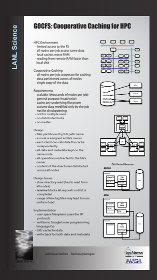


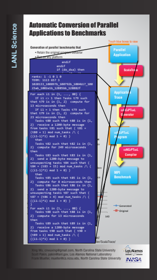






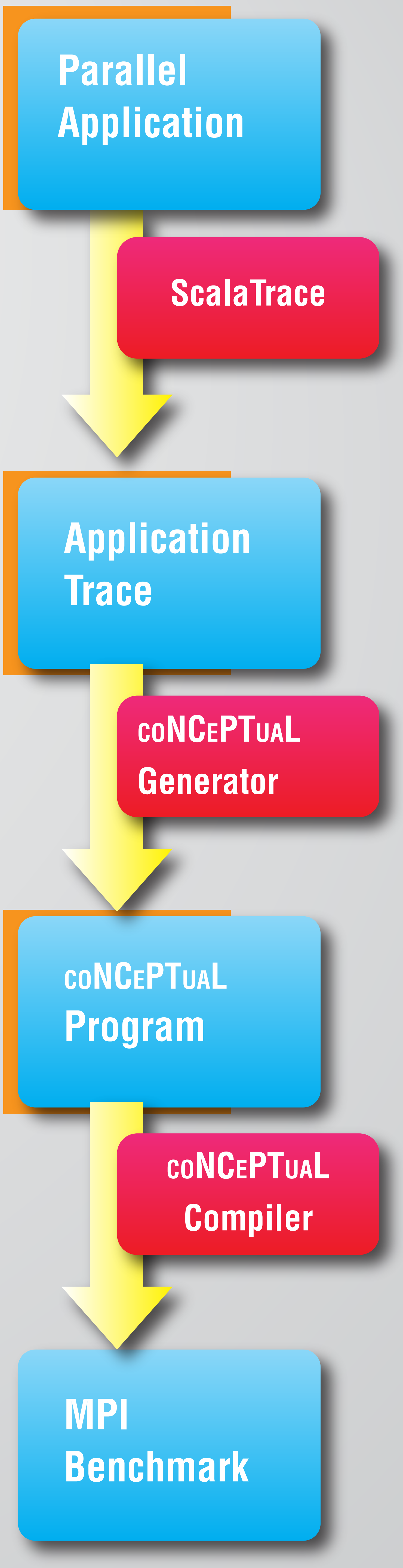






Automatic Conversion of Parallel Applications to Benchmarks

Touch blue boxes to view



Generation of parallel benchmarks that

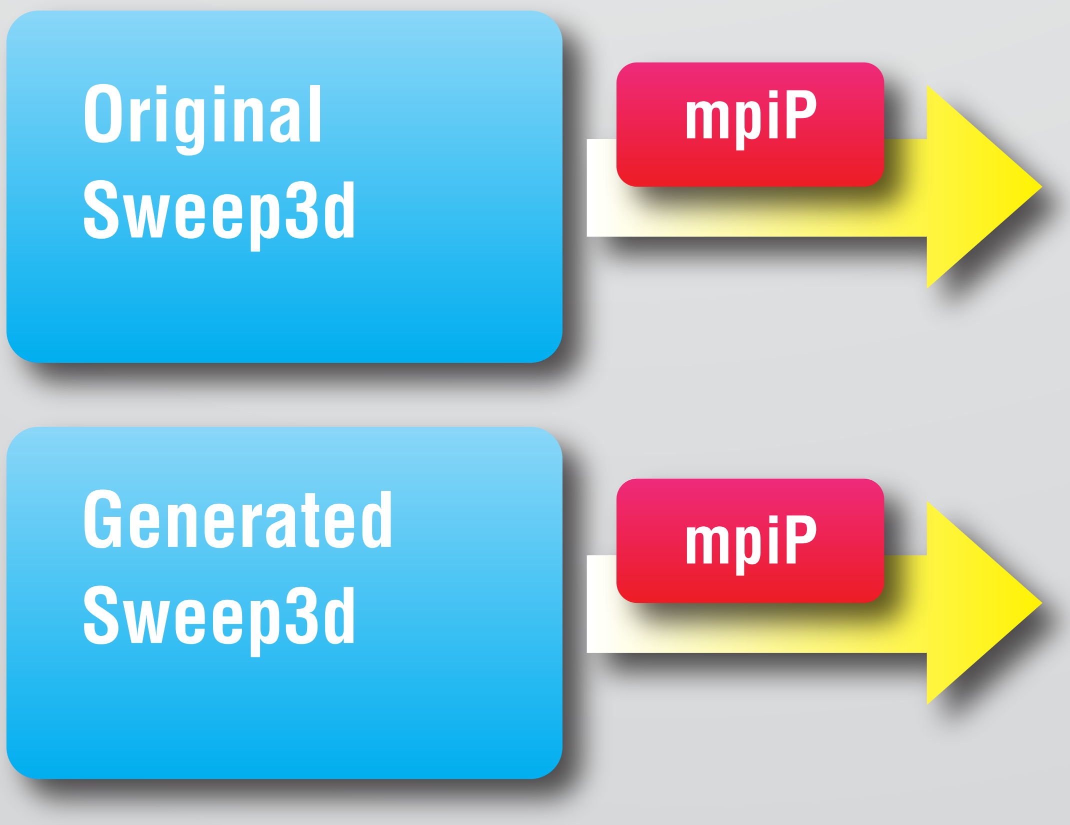
- Retain the original program behavior
- Run on any platform
- Are simple and easy to understand

Methodology: App → Trace → Benchmark

- **ScalaTrace:** a highly scalable parallel application tracing tool
- **coNCePTuaL**: a programming language and compiler for rapid generation of network benchmarks
- **coNCePTuaL Generator:** a tool that generates coNCePTuaL programs from traces obtained from parallel programs

Example: Sweep3d

- Original: 1996 lines (Fortran)
- Generated: 2127 lines (coNCePTuaL); will eventually be significantly smaller
- Verification:



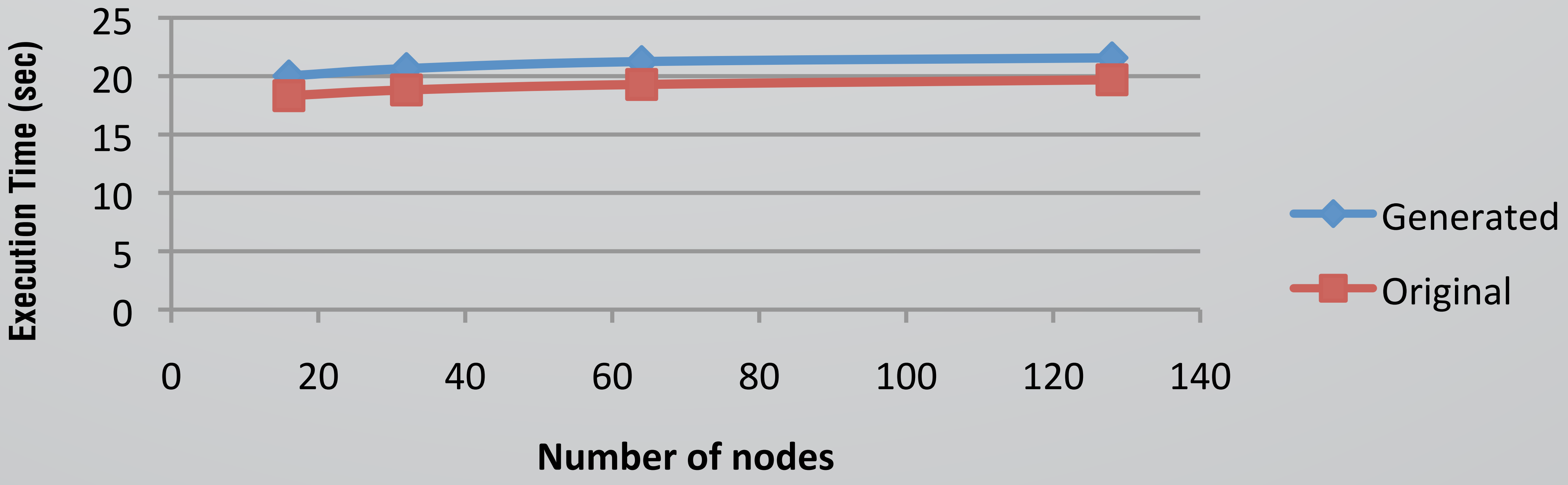
✓ **MPI operations**

153600 Send/Recv
64 Bcast
448 Allreduce
48 Barrier

✓ **Msg volume**

1.84e+08 Send/Recv
2.62e+03 Bcast
2.94e+03 Allreduce

Exp2. Compare execution time



coNCePTuaL: http//conceptual.sourceforge.net/
ScalaTrace: http://moss.csc.ncsu.edu/~mueller/ScalaTrace/

Xing Wu, cnwuxing@gmail.com, North Carolina State University
Scott Pakin, pakin@lanl.gov, Los Alamos National Laboratory
Frank Mueller, mueller@cs.ncsu.edu, North Carolina State University